DOCUMENT RESUME

ED 465 243 EC 308 999

AUTHOR Minnema, Jane; Thurlow, Martha; Bielinski, John

TITLE Test and Measurement Expert Opinions: A Dialogue about Testing Students with Disabilities Out of Level in Large-Scale Assessments. Out-of-Level Testing Report.

INSTITUTION National Center on Educational Outcomes, Minneapolis, MN.; Council of Chief State School Officers, Washington, DC.; National Association of State Directors of Special Education, Alexandria, VA.

SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
REPORT NO NCEO-R-6
PUB DATE 2002-03-00
NOTE 33p.
CONTRACT H324D990058
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; *Adaptive Testing; Criterion Referenced Tests; *Disabilities; *Educational Assessment; *Educational Testing; Educational Trends; Elementary Secondary Education; Evaluation Methods; Evaluation Problems; Focus Groups; Knowledge Level; Norm Referenced Tests; Psychometrics; *Student Evaluation; Test Validity; Testing Accommodations; *Testing Problems; Testing Programs; Validity

ABSTRACT

Two focus groups of test and measurement experts were held to explore the use of out-of-level testing for students with disabilities. The participants (n=17) included state and federal level assessment personnel, test company employees, and university professors. A content analysis of the narrative results indicated that there was no clear consensus in supporting or not supporting out-of-level testing. Instead, focus group participants were able to adopt numerous perspectives on many contentious issues. Four key areas demonstrate salient patterns in the focus group data. First, discussions reflected multiple issues and varying definitions that are reported in the field through testimonial evidence. Second, there was marginal comfort in equating out-of-level test scores back to on-grade level test scores for reporting purposes when the state test was a norm-referenced instrument. This "comfort" decreased when participants discussed criterion-referenced instruments, especially when students were tested more than one level below their assigned grade level. Third, there was general consensus about the need to develop large-scale assessment instruments with broad based content so that more students can be included in the testing program. Finally, opposition to out-of-level testing centered on out-of-level policy concerns rather than psychometric concerns. Appendices include focus group probes. (Contains 13 references.) (CR)

Reproductions supplied by EDRS are the best that can be made from the original document.

Out-of-Level Testing Report 6

ED 465 243

Test and Measurement Expert Opinions: A Dialogue About Testing Students with Disabilities Out of Level in Large-Scale Assessments

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

BEST COPY AVAILABLE

EC 308999

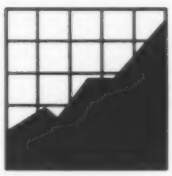
Test and Measurement Expert Opinions: A Dialogue About Testing Students with Disabilities Out of Level in Large-Scale Assessments

Jane Minnema • Martha Thurlow • John Bielinski

March 2002

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission, provided the source is cited as:

Minnema, J., Thurlow, M., & Bielinski, J. (2002). *Test and measurement expert opinions: A dialogue about testing students with disabilities out of level in large-scale assessments* (Out-of-Level Testing Report 6). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Out-of-Level Testing Project supported by a grant (#H324D990058) from the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

NCEO Core Staff

Deb A. Albus
John S. Bielinski
Jane L. Krentz
Kristi K. Liu
Jane E. Minnema
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, Director

Additional copies of this document may be ordered for $15.00 from:

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://education.umn.edu/NCEO

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

This document is available in alternative formats upon request.

Executive Summary

Fourteen states now allow out-of-level testing as a statewide testing option for students with disabilities: Arizona, California, Connecticut, Delaware, Hawaii, Iowa, Louisiana, Mississippi, Oregon, South Carolina, Texas, Utah, Vermont, and West Virginia. Generally, this testing option is based on the belief that matching the level of a test to an instructional level will produce a better measure of a student's true ability level. Still, there are many issues surrounding the psychometric properties of out-level-testing and the accuracy and precision of the resulting test scores. Two focus groups of test and measurement experts familiar with out-of-level testing were held to begin to resolve some of these issues.

A content analysis of the narrative results indicated that there was no clear consensus in supporting or not supporting out-of-level testing for students with disabilities in large-scale assessments. Instead, focus group participants were able to adopt numerous perspectives on the many contentious issues that surround out-of-level testing at the local, state, and federal levels of the educational system. Themes of results did emerge from discussions on the advantages and disadvantages of out-of-level testing. For instance, participants suggested that out-of-level tests could provide a better testing experience for some students, could meet unique assessments needs, might be a fairer approach to testing, and is often favored by parents of students with disabilities. On the other hand, in considering the disadvantages of out-of-level tests, participants were concerned that out-of-level testing is open to misuse, is problematic in reporting test results to the public, and is often put in place by individuals with little assessment literacy.

Four key learnings reflect salient patterns in the focus group data. First, both of the focus group discussions reflected multiple issues and varying definitions that are reported in the field through testimonial evidence. Second, there was marginal "comfort" in equating out-of-level test scores back to on-grade level test scores for reporting purposes when the state test was a norm-referenced instrument. This "comfort" decreased when participants discussed criterion-referenced instruments, especially when students were tested more than one level below their assigned grade level. Third, there was general consensus about the need to develop large-scale assessment instruments with "broad based" content so that more students can be included in the testing program. Finally, the opposition to out-of-level testing that emerged centered on out-of-level testing policy concerns rather than psychometric concerns. These concerns reinforce the need to conduct an experiment that determines the differential results of out-of-level test scores compared to on-grade level test scores.

Overview

First introduced in the 1960s, out-of-level testing was used to measure student academic progress as an indicator of Title I program efficacy. It was reasoned at that time that matching test item content to students' ability levels, rather than their assigned grades, yielded more reliable and valid test results. In other words, if a 5th grade student was reading at a 3rd grade level, a 3rd grade level reading test would be a more precise and accurate measure of the 5th grade student's reading skills. Today, some educators, parents, and policymakers continue to embrace the logical assumption that matching a level of a test to an instructional level will be a better measure of a student's true ability level (Minnema, Thurlow, & Scott, 2001). While the logic seems straightforward, a closer look at a program of out-of-level testing raises two major concerns.

A first concern has to do with the assessment context, which has changed dramatically since the early days of testing students out of level. Out-of-level testing, in its original inception, was used with norm-referenced instruments for which test companies had included common test items across adjacent test levels. In other words, the test items at the ceiling of one level of a test measured the same academic skills as the floor of the adjacent test level. By doing so, test developers created a common measurement scale for a series of test levels that could be administered either below or above a student's assigned grade level. Today, however, some states are electing to test students out of level on criterion-referenced tests used for student and system accountability. Since most criterion-referenced tests are not developed with a common measurement scale for all grade levels of the instrument, testing students out of level is problematic. (See Thurlow and Minnema, 2001 for an extensive discussion of these issues.)

A second concern about testing students with disabilities out of level derives from the psychometric properties of out-of-level test scores. To date, no program of research has clearly delineated the precision and accuracy of out-of-level test scores, or has determined how these psychometric characteristics affect the test results (Bielinski, Thurlow, Minnema, & Scott, 2000). Bielinski et al. (2000) also raised concerns about the precision of out-of-level test scores when the out-of-level scores are equated to in-level test scores in norm-referenced testing. The process of transforming out-of-level test scores to in-level test scores may introduce additional measurement error causing detrimental effects on test score reliability. With respect to accuracy, the literature has yet to demonstrate that out-of-level tests yield more accurate, and therefore more usable, test information for making instructional decisions. These same issues hold true for criterion-referenced tests, especially if the levels of an instrument are not developed with a common scoring scale.

Given the unknown psychometric effects on test score precision and accuracy when students with disabilities are tested out of level, it is difficult to ascertain students' academic progress over time with a high degree of confidence. Further, when out-of-level test information is used

for accountability purposes, especially when making high stakes decisions for students and schools, it is imperative that test data be both accurate and precise. Whether a norm-referenced or a criterion-referenced test, one of the key issues within today's reform-minded assessment context is that the precision and accuracy of out-of-level test scores are questionable (Bielinski et al., 2000).

Taken all together, there is a need to research the psychometric concerns about out-of-level test scores for both norm-referenced and criterion-referenced large-scale assessments. Unfortunately, as is typically true of research on educational policy, the practice of testing students out of level has preceded research on the topic. In fact, past research studies on out-of-level testing seemed to raise more questions than they answered (Minnema, Thurlow, Bielinski, & Scott, 2000). A limited number of research studies conducted throughout the 1970s and 1980s began to parse apart the complex psychometric questions that surround out-of-level testing (Minnema et al., 2000). Even so, none of these studies unconditionally recommended testing students with disabilities out of level (Cleland & Idstein, 1980; Jones, Barnette, & Callahan, 1983; Yoshida, 1976).

Without a solid research base on which to develop sound policy decisions, the practice of testing students with disabilities out of level has evolved within a contentious atmosphere. It has been reported that educators, parents, and state legislators dispute the value and the challenges in testing students out of level at the local, state, and federal levels of educational systems (Minnema et al., 2001). Furthermore, the decision to allow out-of-level testing has often been decided within heated debates among stakeholders who have little knowledge about the precision and accuracy of tests that measure academic progress appropriately (Minnema et al., 2001).

The limited research knowledge to support testing students with disabilities out of level as well as the manner in which policy decisions are made is particularly disturbing since the number of states that allow out-of-level testing as a component of their statewide assessment program has grown rapidly, and may continue to do so (Thurlow & Minnema, 2001). For instance, as of December 2000 there were 12 states that were implementing a program of out-of-level testing in large-scale assessments. Since that time, 14 states (Arizona, California, Connecticut, Delaware, Hawaii, Iowa, Louisiana, Mississippi, South Carolina, Oregon, Texas, Utah, Vermont, and West Virginia) are testing students out of level in statewide tests (Thompson & Thurlow, 2001). Georgia and Alabama are the only states that we know of that have considered and then decided not to use out-of-level testing in their large-scale assessment programs (Jean Cohen, personal communication, June, 7, 2001; Gloria Turner, personal communication, July 19, 2001). North Dakota, a state with a long history of testing students out of level, has recently reversed their decision so that they no longer allow out-of-level testing (Jean Newborg, personal communication, October 19, 2001).

Given the rapid expansion of out-of-level testing, coupled with the psychometric issues that surround testing students out of level, the National Center on Educational Outcomes (NCEO) conducted a study to begin to understand the psychometric concerns about out-of-level testing. This study was designed to gather perceptions and opinions held about out-of-level testing by test and measurement experts. Those results are presented in this report to serve as one perspective on the value and the challenges of testing students with disabilities out of level.

Method

We used an inductive approach to gather narrative data from two focus groups, using the format and procedures recommended by Krueger (1994). The groups were convened during the Council of Chief States School Officers (CCSSO) Large-Scale Assessment Conference in Snowbird, Utah in June, 2000.

The participants (n = 17) included state and federal level assessment personnel, test company employees, and university professors. Specific criteria were used to select our purposive sample from conference attendees. Each participant had an extensive professional or academic background in assessment and testing issues. Prior to agreeing to participate, all participants indicated a familiarity with out-of-level testing and the issues that surround testing students with disabilities out of level in large-scale assessments. Finally, all participants received a copy of the focus group questions a week before the conference to ensure that all participants could participate meaningfully in the focus group conversations.

To begin each focus group, the facilitator read a script that introduced the activity, described the focus group process, defined out-of-level testing, and proposed ground rules for participation. (See Appendix A for a copy of this script.) At this time, each participant received a packet that contained a written definition of out-of-level testing to engender a common understanding among the group. The packets also had two executive summaries of recent NCEO Out-of-Level Testing reports as a thank you for their participation. As an inducement for participation, we provided either lunch or dinner during the focus group session.

Five focus group questions were presented over approximately two hours (see Appendix B for the Focus Group Question Protocol.) A general question was posed first to foster a comfortable atmosphere in which the participants could engage in meaningful dialogue. Participants were asked to answer this question in a round-robin style of participation where each participant spoke in order of seating. The four questions after that contained more specific content, addressing the advantages and disadvantages of out-of-level testing in large-scale assessments and the uses of out-of-level test scores for system and student accountability purposes. These four questions

were answered in a natural give and take style of normal conversation, whereby participants could contribute as they wished. Each focus group conversation was tape recorded for transcription to prepare the data for content analysis.

Results

Our analysis of the narrative data set yielded themes of results; these are presented here for the first two questions posed to the focus group members. The dialogue from the question used to open each focus group was not included in our analysis because the responses were global in content, and was not intended to contribute information to the final results.

Identifying Advantages of Out-of-Level Testing

The qualitative analysis of the responses to the first focus group question that addressed the advantages of testing students with disabilities out of level yielded three strands of results: student-related advantages, teacher-related advantages, and system-related advantages. Themes of narrative findings are presented within each of these strands of results.

Student-Related Advantages

Four themes emerged in the data analysis that fit a category related to students' testing experiences when tested out of level. These themes, which are discussed here, focused on: (1) better testing experience, (2) meets unique assessment needs, (3) fairer approach, and (4) logical approach favored by parents (see Table 1).

Theme 1. Testing students at their instructional level provides a better testing experience for the student.

It seems logical to assume that testing students at their instructional level, even if that level is below the grade level in which they are enrolled, will garner more accurate assessment information. "To me the decision about whether it's appropriate to do out of level testing depends on whether that test is a better alignment to the curricular opportunities and experiences of the student." When test item content is aligned with a student's instructional level, "the value of an out-of-level test is getting some information, which is better than finding out that a kid is in the first percentile or got two points out of 50." In other words, out-of-level testing provides "feedback on whether students are learning what they're being taught. Also, you want to be able to get

Table 1. Focus Group Results for Question 1

Q1. What are the advantages of using out-of-level testing for students with disabilities in large-scale assessments?	
Student-Related Advantages	Theme 1 - Better testing experience Theme 2 - Meets unique assessment needs Theme 3 - Fairer approach to testing Theme 4 - Logical approach favored by parents
Teacher-Related Advantages	Theme 1 - Better information for teachers Theme 2 - Reduced use of test modifications
System-Related Advantages	Theme 1 - Includes more students in testing Theme 2 - Promotes school change

feedback that shows strengths as well as weaknesses or what students can as well as can't do. It's possible that out-of-level testing would help you with that." In turn, the test experience should be a less frustrating experience, and one that promotes a sense of well being for a student. While some of the logical thinking about out-of-level testing may be suspect, there are converging sources of testimonial evidence and narrative data that support the contention that taking a state test out of level may be a better test experience for students than taking the test on grade level.

State education agency personnel have reported that teachers and parents expressed concerns that participating in a state test that is too difficult has negative ramifications for students (Minnema et al., 2001). In fact, the reported reactions tended to be highly charged with emotion; emotional reactions from teachers and parents and emotional reactions of students during the testing situation. This report of an emotional reaction to participating in a state test at the grade level in which a student is enrolled also emerged in our focus group data, as evidenced by the following comment: "As a state assessment director, I didn't have the experience of positive reinforcement. I only had negative experiences from the letters and calls I've received." A later comment by the same participant reflected teachers' frustration about observing their students' test taking experiences. "I've spent two years working on this student to get him some self-esteem and you've just destroyed it. You have just destroyed two years of my work."

Theme 2. Out-of-level testing is an individualized approach to testing that meets students' unique assessment needs.

According to some of the focus group participants, students with disabilities have unique assessment needs that require an individualized approach to testing. Out-of-level testing provides a "customization of testing to the particular student [that] is of tremendous advantage. Basically, you're catering to each students' needs" by administering a test at a student's particular level of functioning. When an assessment is directly measuring academic constructs from a students' curricular level, the test results provide "feedback that shows strengths as well as weaknesses or what students can as well as can't do." Since the level of an out-of-level test is not determined by a student's assigned grade level, the test results "do show what students are capable of, not what they're not capable of. It gives very meaningful feedback on individual students."

Theme 3. Testing students out of level is a fairer approach to assessing students who are instructed at a level lower than their assigned grade levels.

Testing students who are accessing curricular content at a grade level lower than the grade level in which they are enrolled appears to have a high level of face validity. Collecting assessment data on the constructs that are presented to students during their instructional delivery seems logically to be a more valid measure of academic progress. This logic extends to the usefulness of the test scores also. On the surface, information gathered at students' levels of academic functioning should be more usable information for teachers to make sound instructional decisions. Taken all together, an out-of-level test experience is thought to be a more accurate and precise measure of students' skills and knowledge.

Participants in both focus groups indicated that "for the individual student, to get some information about the child, you need to test them at the level at which they're functioning." By doing so, "there seems to be an inherent sense of fairness for the students. It doesn't seem to be productive to be asking them a bunch of questions about material that they've never been exposed to." There also was some sentiment expressed concerning the fairness to the teacher in that for "an 8th grade student who is receiving instruction in the curriculum at the 4th grade level, it's not fair to the teacher who's trying to teach that student to do an assessment on 8th grade material." Further, teachers receive assessment data that are more applicable to their instructional program. "An out-of-level assessment would give me direction in knowing what kinds of skill deficiencies could best be addressed in order to get the student ready to get to the point, if ever, to take the on-grade level test."

Theme 4. Parents prefer out-of-level testing because it seems to be a logical solution for testing students with disabilities in large-scale assessment programs.

While parents were not referred to frequently in the dialogue of either focus group, one participant in particular spoke to the preferences of parents of students with disabilities in terms of reporting test score data. Of the parents with whom he had contact, they thought that "if a 12 year old student who's disabled is being instructed as if [he or she was] 10 or 11 years old, it would only be appropriate to report their scores with the 10 or 11 year old scores." According to the one participant, parents also seemed to prefer "interpreting the scores in terms of student and system accountability" by comparing them to students who may be younger but are functioning at a similar academic level as their child.

Teacher-Related Advantages

Further analysis of the narrative responses to the first focus group question illuminated two themes of results that pertain to teachers who test students out of level. These themes focused on: (1) better information for teachers, and (2) reduced use of modifications (see Table 1).

Theme 1. Teachers have more valid and meaningful test results to use for instructional decisions when students with disabilities are tested out of level.

Some participants indicated that they think out-of-level test scores are "the most valid measure of a what a student is learning. It [out-of-level test] represents the level that the student is receiving instruction on. It makes perfect sense to me." Test items that gather information at the level at which a student is learning seem to logically inform content area decisions for instruction. The level "the student is being instructed on seems the most appropriate point at which the assessment should go on." However, "the issue is the appropriateness of the test and matching the appropriateness of what is going on instructionally." If there is a mismatch between test item content and a student's level of academic functioning, "the purpose of using your assessment system to develop individualized instruction" is not an option.

Theme 2. Out-of-level tests eliminate the need to modify a grade-level test for some students with disabilities.

In some states, teachers and other Individualized Education Program (IEP) team members have the latitude to "modify an assessment. The assumption is that teachers will make modifications based on what they know about the student's abilities and what they know about the curriculum that's being taught." In this case, teachers generally present the grade-level version of a state test, but amend the passing score on an individual basis. However, "sometimes it gets down to

a ridiculous level where they're getting about 25% of the questions correct." The test results in this case provide little usable information for classroom programming. "An out-of-level test could at least put those kids back onto a measurement tool that would give some accurate information about what they were able to do."

System-Related Advantages

The concluding two themes of results that identify advantages of out-of-level testing focus on educational systems in general. The themes address: (1) including more students, and (2) promoting change (see Table 1).

Theme 1. Using out-of-level testing in statewide tests is a means to include more students with disabilities in large-scale assessments and accountability programs.

The line of thinking that emerged in response to the first focus group question affirmed the need to implement inclusive testing programs that support full participation for as many students as possible. Participants generally agreed that out-of-level testing programs include more students at a level at which they can participate. "The advantage is you get to include special education in the assessment program where the only option may be to test them out of level." However, both focus group conversations qualified this advantage by saying, "If the option is to exempt them from the on-grade assessment or test them out of level, my preference would be to test them out of level." One participant stated further, "I'd want to know how the information is going to be reported so it's clear that the student is being tested out of level. I'd also want to hear what source of support people are getting so they don't misinterpret the results."

Theme 2. The implementation of an out-of-level testing policy can promote school system change.

A discussion arose about the purpose of putting educational policy in place that included specific aims such as instructional decision making or student accountability. One participant suggested a less apparent reason for adopting specific assessment policy that is an important consideration in understanding the rationale for out-of-level testing: The goal of implementing a certain assessment policy may be to put in place a new assessment program. However, an ancillary purpose is to ultimately create more appropriate learning environments for students with disabilities. By implementing an out-of-level testing policy, more students with disabilities are included in assessment and accountability programs. Classroom teachers, who use the test data to make instructional decisions for students who had previously been excluded from state tests, are promoting system change by creating new learning options for students with disabilities.

Using an out-of-level testing program to promote school system change is evidenced by the following comment, "There's another purpose, that is to put in place certain policy incentives for behavior that we [policymakers] want." Changes in educators' "behavior" can "get students with disabilities into different environments for instruction." The systemic change may occur slowly with new decisions made for one student at a time, but the intent of the policy is to promote those discussions among educators that can ultimately make positive changes for groups of students with disabilities.

In responding to the second focus group question about the disadvantages of testing students out of level, both focus groups reflected the unresolved and contentious issues that surround out-of-level testing at the federal, state, and local levels of education (see Minnema et al., 2001 for a more in depth discussion of these issues.) In fact, both focus groups identified disadvantages during the portion of the dialogue that was structured to concentrate on the advantages of testing students out of level. In other words, some of the participants from both focus groups identified advantages by providing a disadvantage as a caveat. The content analysis of these conversations yielded five themes of narrative results that clustered into two categories: system-level disadvantages and student-level disadvantages.

Identifying Disadvantages of Out-of-Level Testing

System-Level Disadvantages

Our analysis yielded three themes of results that point to disadvantages of testing students out of level, where the effects of the testing operate at the system level of a school district or state education agency. The three themes focused on: (1) openness to misuse, (2) problematic reporting, and (3) policy set by individuals with little assessment literacy (see Table 2).

Theme 1. Out-of-level testing programs are open to misuse.

With the current emphasis on improved academic performance for both students and school systems, states are looking for ways to demonstrate progress over time. When the stakes are high for students, educators, or school systems, it is tempting to exclude either low performing students from statewide testing or to drop their test scores from the aggregated reporting for accountability purposes. In most states that allow out-of-level testing, the lowest performing students tend to be students with disabilities.

The participants in our focus groups confirmed these testimonies by identifying three ways in which out-of-level testing can be misused. First, "If the school is being held accountable, it's

Table 2. Focus Group Results for Question 2

Q2. What are the disadvantages of using out-of-level testing for students with disabilities in large-scale assessments?	
System-Level Disadvantages	Theme 1 - Openness to misuse Theme 2 - Problematic reporting Theme 3 - Policy set by individuals with little assessment literacy
Student-Level Disadvantages	Theme 1 - Invalid test results Theme 2 - Negative effects on classroom instruction Theme 3 - Differential negative effects for some student subgroups

always possible that somebody's going to want to take a child and put him somewhere where he's going to show the best performance." Second, after testing when the scores are submitted to the testing company contracted to analyze a state's large-scale assessment results, "They [out-of-level test scores] are just removed." Third, misuse of out-of-level testing programs can also occur at the point of selecting students for testing below grade level. "What happened was somebody at the school, in this one particular state said, 'Forget it. I'm not giving this group of students the test that they're supposed to get. I will give them this other test.' And that's what they did."

Couched within the multiple ways that out-of-level testing can be misused was an underlying assumption that states were at least attempting to include more students with disabilities in state or district accountability indices by testing them out of level. This assumption was revealed in the following comment, "I've heard this repeatedly from instructional people, you (SEAs) disenfranchise those kids and you give school systems the opportunity to disenfranchise those kids. By at least including them in the assessment program and figuring out how to deal with the validity of reporting issues, you keep the pressure on school systems to make sure they're paying attention to those kids."

Theme 2. Reporting out-of-level test scores to the public is problematic.

Some participants questioned the validity of out-of-level test scores. In fact, one participant commented, "As practiced most of the time, the out-of-level testing doesn't provide, or quite frequently does not provide, valid information on the construct of interest." Assuming this to be true, "The question then becomes what do you do with the [test] data." Some of the discussion differentiated reporting practices for norm-referenced and criterion-referenced statewide tests.

If a norm-referenced test is administered out of level and appropriate equating procedures are used to transform a lower grade test score to the grade level in which a student is enrolled, the decision to report a test score on grade level can be done with some confidence. Test companies conduct equating studies to develop normative data thereby linking various grade level test scores on a common scoring scale. "In the case of a multi-level test that's vertically scaled, testing a 3rd grader using a 2nd grade test and the norms of a scoring table ... putting that out-of-level test score back into the 3rd grade level" is not problematic. Believing this to be true, states can report an out-of-level test score with the grade level scores in which the student is enrolled. For this situation, however, participants did caution that "there's some boundary in the out-of-grade level [testing programs] where I think that scaling would be more comfortable than others." When an NRT is developed so that adjacent grade levels contain overlapping test items (e.g., the ceiling of a 4th grade test would use test items that are similar to the floor of a 5th grade test), participants felt more comfortable reporting an out-of-level test score on-grade level if the gap between the test grade level and the student's assigned grade level was limited to a few grade levels. However, when there are "giant differences in the grade levels" of a test level and the student's grade level of enrollment, participants expressed discomfort in combining out-of-level test scores with on-grade level test scores for reporting purposes.

Additional apprehension emerged in the discussion around reporting out-of-level test scores for a criterion-referenced statewide test. Large-scale assessment programs that use criterion-referenced instruments do so to measure groups of students' progress toward achieving grade level content standards. In this way, states can monitor academic progress over time by grade levels or by student subgroups. For instance, when state test scores are disaggregated and reported for subgroups of students such as students with disabilities, it is assumed that the reported results are aligned with these students' curriculum. While the participants in these focus groups supported reporting aggregate and disaggregated test data by grade levels, they questioned the practice of combining out-of-level test scores with on-grade level test scores when the results measure certain specifications on a continuum of academic skills and knowledge. "If it was the case that you were actually trying to measure different things at different levels," as criterion-referenced tests do, "then we don't have the comparability of the scaling." Test data that are reported in aggregate by combining test results from two different grade levels is not mathematically sensible. Since these combined test results represent academic progress toward different criteria, the combined results are not a pure measure of either the out-of-level or on-grade level academic progress. "From a standard-based content point of view, I have a terrible problem with that." In other words, when an out-of-level test is used to make evaluative decisions such as demonstrating academic progress toward grade-level standards, these participants registered "real practical limitations" in developing "adjacent level tests that would be built to measure functional levels" when the content should be fairly different between the two [grade] levels tested."

Theme 3. The decision to allow out-of-level testing is frequently made by policymakers who have little assessment literacy.

If out-of-level testing was mandated by a legal body such as a state legislature, additional assessment and accountability problems ensue. Some participants, particularly those who were involved in state level decision making about out-of-level testing, reported that they "didn't experience positive reinforcement" from practitioners and parents who advocated for excluding students with disabilities from participating in regular state assessments. State legislators also received political pressure from their constituency. In this case, well meaning advocates set up a situation for policymakers to think about assessing students in special education differently from students in general education. "The problem is that the legislation that we have in place really deals with labels." In other words, policymakers discussed policy options in terms of general education and special education rather than in terms of tests that "measure the same construct" or "the measurement being the same" for both groups of students. Participants did acknowledge that even though the solution to including students with disabilities in large-scale assessment programs by using out-of-level testing is less than satisfactory, "[when] you require it to include disabled students in an assessment program, if your only option is to test them out of level, then maybe it's the only option that you have. Then you have to do it."

Student-Level Disadvantages

We identified three themes related to student-level disadvantages. The effects of these disadvantages directly impact individual students' test performance or academic progress. They are: (1) invalid test results, (2) negative effects on classroom instruction, and (3) differential negative effects for some subgroups.

Theme 1. Out-of-level tests yield invalid test results.

A primary concern among some of the participants was the integrity of the test score from an out-of-level test, as indicated by the following comment, "As practiced most of the time, the out-of-level testing doesn't provide ... valid information on the construct of interest." The validity of out-of-level tests was questioned in two ways: first, by the psychometric properties of the instruments used for testing students out of level, and second by the item content in those tests.

Since many states use criterion-referenced tests for large-scale assessment programs, participants pointed out the difficulties in using this type of instrument for testing below grade level. "If the tests are measuring different things (e.g., algebra at the 8th grade level and basic math skills at the 5th grade level) then we don't have the comparability of scaling." However, in the case of norm-referenced testing, there are scaling methods that can establish a common scale to equate multiple test score levels translating below level test scores to on level test scores. From a

psychometric perspective, some participants indicated that in this testing situation, they would consider the out-of-level test scores to be valid test results. However, if there were "giant differences both in the grade levels and the specification differences, the validity of the out-of-level test score would be problematic." For both CRTs and NRTs used for testing students out of level, "if this test is measuring algebra and this test is measuring basic skills, then no, I can't make standards-based conclusions based on the scaling."

A few participants spoke to the issue of test validity when "an assessment [is in] alignment with the children's experiences. ... it's appropriate to do out-of-level testing depending on whether that test is a better alignment to the curricular opportunities and experiences of the student." Even though stating that out-of-level testing may be technically appropriate for some assessment situations, participants followed this viewpoint with two caveats.

First, there was concern in that "we were looking to test the same [construct] across all of the age spans but we were very sensitive to presenting it in a context that would be amenable and familiar to kids in that level. So I think ... if you've got something that's really going to work with one age group area, it will be a disadvantage to others or render it less accessible." Some tests were developed with "the level of language ... similar across all the grade spans" for testing below grade level at multiple ability levels. Test developers "were looking at the 3rd and 4th grade pieces differently than the 7th and 8th grade pieces. We had to clearly look for something that kids with very little language could access but that they wouldn't think was too babyish." An age-inappropriate instrument could affect students' motivation so that assessments are not taken seriously, resulting in test scores that are not valid representations of what students know.

A second caveat concerned the content of the testing instrument. Some participants suggested that an assessment needed to be matched to the student's assessment needs. Since "state testing programs have gone the way of having an elementary, a middle, and a high school level, your choices aren't as graded as you might need them to be." The test's validity is partially dependent on how well the test items "align with where they're receiving instruction. The expectation is that they will be making gains" when their academic progress is "referenced against the curriculum that is appropriate for the level that they're receiving instruction."

Theme 2. Out-of-level testing may have negative effects on classroom instruction.

Speaking from a policy perspective, participants commented on how out-of-level testing "removes some of the policy incentive to make sure to the extent possible that students are moved into challenging curriculum." There is a concern that has circulated within local, state, and federal educational agencies that questions how well teachers can maintain high learning expectations for those students with disabilities who are tested out of level. If students do not receive on-grade level instruction, they may not be provided the opportunity to learn grade-

level standards. There was some agreement among a few participants that presenting on-level tests would introduce "a little bit more frustration for students in this nation because they're going to be exposed to high level content." However, participants continued to suggest that a major concern about testing students with disabilities below grade level was the possibility that a student would not receive grade-level, standards-based instruction. "I just think we have to be really careful with the rules and all of that to make sure that children aren't being tested inappropriately at a lower grade level and then stuck in a dead-end curriculum to boot."

Theme 3. Testing students out of level may have deleterious effects on certain subgroups of a school district's student population.

A topic in the conversation during both focus groups reflected concerns about "disenfranchising" certain students from the benefits of school improvement plans. Generally speaking, out-of-level testing in most states is reserved for testing students with disabilities. However, when groups of students are "set aside" by a different policy from that applied to students in mainstream education, it is likely that these students will be excluded from regular statewide assessments. The result is that "it will disenfranchise [students with disabilities] from instructional decisions that are made on their behalf." Well-meaning educators or parents may select a student for an out-of-level test assuming that their decision will promote better test performance and ultimately, improved educational results. When students are not part of the regular assessment program, "the question is how do we improve [the educational system], who do we improve the delivery [of instruction] to ... if they're not part of the denominator, they're not a part of the solution."

A few participants raised another tangential issue that looks at a different subgroup of students who do not have equitable testing options either. In referring to low performing students who do not receive special education services, one participant commented, "You have some disabled students for whom special testing requirements are necessary in terms of the appropriateness of the level of testing. But you have a lot of non-disabled students who are also equally disadvantaged in terms of their educational setting and structure. You don't test them out of level in the large-scale assessment." In other words, assuming that out-of-level testing appropriately measures some students academic progress, an assessment program that does not provide equitable testing options for all students yields test data that do not support equitable school improvements for all students.

One other set of ideas that emerged from the conversation focused on disenfranchising groups of students from assessment programs. In terms of selecting students with disabilities for an out-of-level test, participants indicated that "there are plenty of low functioning non-special education or non-LEP kids [for whom] we can't show anything close to what they're doing in their nominal grade level assessment." Since only some select groups of students can meet the

out-of-level testing criteria in most states that allow out-of-level testing, "Whoever's making the decisions about whether to test kids on-grade level or out-of-level ... need to be really clear about what the rules are and the rationales." In other words, it is essential to "have good rationales about making [out-of-level testing] decisions."

Using Out-of-Level Testing for Student and System Accountability

The third and fourth focus group questions asked about the appropriate uses of out-of-level testing for student and system accountability programs. Since accountability is defined differently across all states, we provided a "typical" definition for participants to use in framing their responses. For the purposes of this report and the facilitation of our focus groups, we defined accountability as an individual or group of individuals who take responsibility for the performance of students on achievement measures (NCEO, 2001). Student accountability assigns responsibility to individual students who demonstrate academic progress in meeting state content standards by participating in large-scale assessment programs. The second type of accountability, system accountability, holds an educational system, or individuals within the system, responsible for demonstrating improved academic results.

Even with these common definitions, it was difficult for the participants to adhere to understandings that were outside of their personal frames of reference. Our data analysis revealed that most participants couched their responses within the accountability contexts of their own professional experiences. One of the functions of a focus group process is to gather a variety of individual perspectives to amass common strands of information that answer a particular question. However, neither our data analysis nor the summaries provided at the end of each focus group revealed dominant themes of results. Thus, we decided not to treat the responses to the final two focus group questions separately for determining uses for out-of-level testing in each of the two types of accountability systems. Instead, we combined the discussions about student and system accountability from both focus groups into a composite data set, and considered the data set holistically.

Because our process to determine the results for these two questions differed from the approach to analyzing the results from the first two focus group questions, we chose a different format for presenting the interpretation of these data. Instead of presenting themes of results that depict ways that out-of-level testing can be used appropriately for student and system accountability, we identified topics of conversation that appeared to be focal points within each groups' conversations. These focal points emerged as well-developed lines of discussion that point to four specific accountability issues that are important considerations when using out-of-level testing results for accountability purposes (see Table 3).

Table 3. Results Pertaining to Student and System Accountability Issues

Prominent lines of focus group discussion
Issue 1 - Using out-of-level test scores for accountability purposes promotes greater attention to student achievement at the lower end of the continuum.
Issue 2 - Out-of-level testing is better suited for student accountability programs than system accountability programs.
Issue 3 - States differ in how student and system accountability are distinguished.
Issue 4 - Selecting students appropriately for out-of-level tests is linked to the results of system accountability programs.

Issue 1. Using out-of-level test scores for accountability purposes promotes greater attention to student achievement at the lower end of the continuum.

As a positive consequence of testing students with disabilities in large-scale assessments, the line of conversation followed up on an earlier comment: "... when we first started this conversation, one of the things that somebody said is that we don't tend to test these kids [students with disabilities]. In doing so, however, "The positive question is that there are a bunch of kids [who] are left out of the accountability system entirely." Not only is this exclusionary practice discouraged by Title I regulations, but according to one former state testing director, "I heard from their parents, 'Why don't you hold the schools accountable for teaching my children something?'" While "you don't necessarily need to do out-of-level testing to include these children" in a statewide assessment program, it is at least an approach that includes more students with disabilities in the testing. The results may then "prove to the teacher that the child does have more capacity and more ability than he or she was thinking originally." Along a similar line of thinking, one participant noted that just having the conversation about including more low achieving students has value. "The advantage of it is, or of at least having the discussion ... is that it opens up the conversation about greater expectations and instruction of kids who some people think should be tested out of level." These ideas then "open up the conversations to why their instruction is different and if it really should be. And that's a conversation worth having even if it means veins popping out of people's necks."

Issue 2. Out-of-level testing is better suited for student accountability programs than system accountability programs.

Both groups of participants raised concerns about using out-of-level test scores for system accountability purposes. This line of conversation opened with the following comment: "I see that it [out-of-level testing] can be used appropriately if your focus is on expectations that are

tightly aligned with curriculum that's being taught essentially to the same standards regardless of where you're at." If there is a match between test item content, our results suggest that out-of-level testing may be useful for monitoring student academic progress over time. Several comments focused on the uncertainty of curricular alignment to test items, especially when norm-referenced instruments are used for statewide testing. Generally speaking, these participants identified a role for out-of-level testing when teachers needed to make instructional decisions for individual students.

The concern about out-of-level test score use arose when addressing its appropriateness for system accountability purposes. Some participants suggested that if a student was tested below the grade level in which they were enrolled, a zero should be entered into the accountability index within the aggregated test data for the students' assigned grade level. In other words, if a student's academic progress in achieving a set of standards that are intended for a lower grade level than the grade in which a student is enrolled, the state test cannot measure progress toward grade level standards. Entering a zero in the accountability index for this student indicates that the student has made no progress toward meeting grade level content standards. However, for "a person who would want to interpret the results, knowing a student got a zero on a test of content that that student was never exposed to is not very instructionally useful to an instructional planner." The flip side of this argument asserts that out-of-level test scores that are reported as zeros do not denote a student's academic progress fairly. This issue is especially confusing in that the student has not made zero progress toward content standards – just no progress toward the standards for the grade level in which the student is enrolled. Because of this, some participants thought that the student should receive some credit for progressing toward a set of content standards.

The point at which system leaders interpret test score data for entire grades to make system level decisions about improving instructional delivery in a particular content area is especially problematic. If test score data combine out-of-level test scores with in-level test scores, the results represent progress toward various sets of standards that were designed for different grade levels. In this case, system level decisions to improve instruction for specific grade levels are based on test data that reflect several sets of grade level standards. For a classroom teacher who makes instructional decisions, "knowing that the student is doing marginally well or even good, or poor, on material that they are covering, tells me something about the instructional program's success and their achievement and how they are progressing in the instructional program. That's the place where it makes sense to me." Other participants agreed with this assertion that student accountability "is the place where I feel the strongest that out-of-level testing does have promise."

Issue 3. States differ in how student and system accountability are distinguished.

One participant described the crux of this issue by saying, "It's hard for me in the situation that we're in across the country now to think about a system where student accountability is totally separate from system accountability. That doesn't happen anymore. Or at least I don't know where it is." Other responses to the questions about accountability indicated that the need for out-of-level testing was "going to depend on the kind of accountability system." One participant summed up the lack of consistency across states by saying, "Let me just postulate two very different systems and then there are a million variants of all of this. One is the percent of students reaching a certain level. That is all you care about – that percent. If that's your model, you don't need out-of-level testing. An alternative model might be the average score or the percent of people at a lot of different places." In this case, out-of-level testing could provide information for those students whose scores fall close to either the floor or ceiling of a given test level. The participants in one focus group engaged in a lengthy dialogue about this second accountability model. We find these data to be an important part of the on-going conversation about out-of-level testing.

For some system accountability models, the procedures used to calculate the statistics can mask the performance of certain groups of students. For instance, it is possible to "get a floor effect or a ceiling effect," which means that the resulting statistics may not represent some students' actual scores. In this way, "you're going to not see what's really going on. So there could be improvement but your measurement instrument doesn't allow it. Or there could be scores going down and you can't find it because again the floor and ceiling get in the way." Out-of-level testing might be more sensitive to changes in scores that approach either the floor or ceiling of a particular instrument.

As a final example that is indicative of the wide variability across states in structuring accountability programs, another participant described a third possibility. "There's a third possibility which is you have an even more subtle model where you're looking at gains as the measure of student accountability. You've appropriately identified that this student is really performing at a certain level that's going to be different from the other students. So the instruction is in fact tailored to the level that the students have. Now you want to see [whether] they have gained relative to where they were." In this case, some participants indicated that an out-of-level test could more appropriately measure the amount of progress relative to that student's own rate of striving to meet content standards.

Our participants also noted variations in how different states hold different people responsible for demonstrating academic progress. Along a similar line of thinking, some participants indicated mixed opinions about "what level of system accountability you are worried about. If you're worried about individual teachers, then not taking into account what they have to start with seems somehow inappropriate." Another participant countered, "But the flip side is that you get

accused of lowering your expectations" for all students if some students are tested below their assigned grade level. Throughout this discussion, as noted before, the wide variability in the structure of states' accountability systems made it difficult for these participants to speak directly to specific uses of out-of-level testing for either student or system accountability purposes.

Issue 4. Selecting students appropriately for an out-of-level test is linked to the results of system accountability programs.

As part of the discussion that focused on system accountability, some participants raised concerns about using "performance levels for judging the quality of the school program being offered to the students. Either with on-level or out-of-level testing, chances are that you aren't going to measure the quality of program offered to students with disabilities because you are going to be in a range that just doesn't cross that threshold. There are certain types of accountability structures that could be more sensitive to [program efficacy] that few states are actually using at this stage." The underlying assumption here seems to be that students with disabilities are striving to meet a different set of standards from their same-age peers. This is an assumption that may not be appropriate for meeting the academic needs of most students with disabilities.

To further this line of thinking, another participant initiated a conversation topic that extended the groups' attention to how students with disabilities are selected for out-of-level tests. The concern centered on "how to figure out the way to make the state part of how you decide you're going to improve reading." Making sound decisions at the system level to improve instructional practices, and in turn program efficacy, hinges on the test results for all students. When states' out-of-level testing policies specify that only students with disabilities can participate in state tests that are out of level, students with low academic achievement but not identified disability are not eligible for out-of-level tests. "There are plenty of low functioning, non-special education, non-LEP kids. We can't show anything close to what they're doing in their nominal grade level assessment." In turn, these students are also at risk for not receiving the benefits of school improvement planning.

Since accountability decisions rest on the interpretation of students' test scores, participants cautioned that, "Whoever's making the decisions about whether to test kids on grade level or out of level, they need to be really clear about what the rules are and the rationales." To best meet all students' assessment needs, one participant summed up what seemed to be the group's sentiment by saying, "I just think that we have to be really careful with the [selection] rules to make sure that children aren't being tested inappropriately at a lower grade level."

Discussion

Our discussion of the focus group results is framed by four key learnings that are important considerations for decision-makers whose states allow out-of-level testing in large-scale assessment programs. Each key learning reflects a salient pattern in our narrative data.

First, both of our focus group discussions reflect multiple issues and varying definitions that are reported in the field through testimonial evidence. Both focus groups spoke to the complexities that surround out-of-level testing programs. However, little consensus emerged in our data that pointed to specific ideas for states to consider when allowing out-of-level testing. For instance, there was discussion around the allowable gap between the grade level of an out-of-level test and a student's grade level placement. No one suggested a specific number of levels below grade level that would be appropriate for an out-of-level test; as another example, both focus groups discussed the problems in defining the purpose of the test be it for student accountability or system accountability. Yet, there were no specifics for states to use to better sort out the issues that pertain to using out-of-level test scores for accountability purposes.

Second, our results suggested that there was marginal "comfort" in equating out-of-level test scores back to on-grade level test scores for reporting purposes when the state test was a norm-referenced instrument. There was also some support for using those test scores for instructional decisions when student accountability was in question. However, there was no reference in either focus group discussion that acknowledged the common concern about using an NRT to test students with disabilities, a group not generally included in normative samples (The Psychological Corporation, 1993). For instance, the equating studies that test companies conduct to formulate the normative data used to transform out-of-level test scores to in-level test scores traditionally under-represent students with disabilities.

One focus group conversation pursued the issues of measuring academic gain over time when the assessment instrument may not be sensitive to smaller increments of progress. In addition, some participants were concerned that relatively large numeric ranges within proficiency levels for reporting test performance might be too broad to demonstrate progress, especially for students with disabilities whose academic progress may be slower than their same-age peers. However, it can be shown that those test scores that fall in either tail of a normal distribution curve are generally biased and saturated with measurement error (Kim & Nicewander, 1993). Since it is fairly safe to assume that some students with disabilities score within the lower range of performance, the validity and reliability of their test scores is suspect. Complicating the problem further, states are attempting to demonstrate progress within the lower proficiency level of performance on a statewide test. To do so, a proficiency level that is neither valid nor reliable is segmented into increments that are also neither valid nor reliable. While progress can be recorded for individual students who perform at the lowest proficiency level, the problem is not eliminated

since the segmented ranges of test scores within the lowest proficiency level of performance remain invalid and unreliable. Our focus group results addressed the many problems in using out-of-level test scores for accountability purposes, but did not arrive at any concrete ideas about this critical issue.

Third, there was general consensus about the need to develop large-scale assessment instruments that are "broader based tests." States typically take a "fix the assessment program" approach when certain subgroups of students are known to be participating at relatively low rates. In other words, states tend to add another instrument to the battery of statewide tests that will be a more inclusive measure of some students' skills and knowledge. There seems to be little impetus in the field to develop a new assessment program that is universally designed for all students in a school district. Some state directors of assessment have indicated that they do not have the resources available to revamp an assessment system, especially when reconstructing the testing program would be prohibitively expensive (Minnema et al., 2001). However, it has been noted that the process involved in developing new assessment systems that are broad-based in test content would not be as expensive as some state personnel fear (NCEO, 2001). Our focus group participants identified the need for more inclusive state tests in conversation, but again offered no possible solutions for states to consider as they grapple with a large-scale assessment program that is not appropriate for all students.

Finally, the opposition to out-of-level testing that emerged in these focus group results centered on out-of-level policy concerns rather than psychometric concerns. These participants raised various concerns about developing and implementing out-of-level policy that was problematic for states at both the state and local levels of educational systems. For instance, there was discussion about the appropriate identification of students whose assessment needs could be best met by an out-of-level test. Participants mostly agreed that the rationale for selecting a student for an out-of-level test needed to be derived from a sound decision-making process that was well documented. High-quality decision-making by a student's IEP team requires concrete criteria that guide the decision to test a student out of level. These criteria need to be directly linked to a student's past assessment performance and predictive of future improved test performance. While our participants enumerated the necessary policy pieces that could support appropriate student selection for out-of-level tests, they did not offer specific information for developing the content of these selection criteria.

There seemed to be general agreement within both focus groups about the mathematical explanation for converting out-of-level test scores to in-level test scores. Most of the concern seemed to arise within the context of using a criterion-referenced instrument for out-of-level testing. Disagreement emerged in one focus group about the issue of test and instructional alignment. Best assessment practices recommend such alignment, but our participants questioned whether test item content was actually aligned with student's curricular content. This concern

also surfaces within the field as testimonial evidence (Minnema et al., 2001). While, our focus group participants identified the issue, they did not put forward any suggestions for resolving this measurement problem.

Overall, the focus groups tended to dialogue in generalities rather than in specifics that might guide out-of-level testing policy development and implementation. To explain these patterns in our data, we look to the data-based information that exists in the current out-of-level testing literature. To date, no study has delineated and explained the psychometric properties of out-of-level tests. Given that context, it is probably understandable that specific recommendations did not emerge from the conversation among this group of test and measurement experts, despite their obvious characteristics of being knowledgeable about out-of-level testing.

Study Constraints

This study is an important step forward in understanding the issues that surround out-of-level testing. It describes the perspectives of a group of stakeholders that has not been previously studied. Even so, there are four aspects to our research design that constrain the interpretation of our focus group data. Two of these constraints are directly related to our sample while the remaining two constraints pertain to the process of conducting our focus groups.

First, our purposive sample was limited to only those test and measurement experts who attended CCSSO's Large-Scale Assessment Conference in 2000. Because of this, our participants did not have similar levels of knowledge and experience with out-of-level testing. We did use specific criteria to select our participants, so that we selected only those participants who indicated that they had enough familiarity with out-of-level testing to be able to comfortably participate in a focus group. Some potential participants declined to participate because of their lack of familiarity with testing students with disabilities out of level. We did, however, rely solely on self-reported familiarity with out-of-level testing to select our participants.

A second constraint was that we were unable to balance our sample by participant characteristics as usually is recommended to avoid biasing the focus group results. Our participants, while representing a variety of employment settings where test and measurement expertise is required, may have entered the focus group dialogue with previous biases about out-of-level testing. In addition, our sample was further restricted by a lack of geographic balance with the U.S., although it is unknown whether there are regional differences in the perceptions and opinions about out-of-level testing that could have biased our results.

Other constraints arise from the process we used. It is generally recommended that a researcher facilitate focus groups to the point at which the information gleaned from the process becomes redundant. This was not feasible because we could not schedule more than two focus groups

within the conference schedule. Thus, our data did not reach a point where we obtained reoccurring themes of results for each theme identified in our content analysis. However, the data do reveal that the participants were not able to produce new ideas on particular topics when requested to do so during the transitions between focus group questions. Finally, the conference schedule also restricted the amount of time within which our focus groups could be conducted. Each focus group was at least 90 minutes or longer in length, but probably could have used additional time to balance the amount of dialogue allotted for each focus group question.

Concluding Remarks

Rather surprisingly, our focus group data did not reveal camps of opposing points of view on testing students with disabilities out of level in large-scale assessments. A few participants did appear to readily identify advantages for out-of-level testing while others appeared more reticent to do so. For the most part, our data revealed multiple perspectives that did not clearly delineate the pros and cons of testing students with disabilities out of level. Our focus group participants tended to be able to speak to both sides of the issue without reflecting the contentiousness that surrounds out-of-level testing in practice.

The absence of strong opinions in our focus group data may be in part due to the lack of extensive research on this approach to testing. Beginning to parse apart the psychometric properties of out-of-level tests is an important first step toward understanding the effects that out-of-level testing has on students with disabilities. This focus group study was a first step in that direction. The next step is to conduct an experiment that determines the differential results, including related factors, in testing students out of level and on-grade level. Once the psychometric issues that surround out-of-level testing are better understood, the research can move toward developing guidelines for decision makers to use in developing and implementing out-of-level testing policy. This type of data-based information better informs state and local decisions about out-of-level testing programs, which in turn can improve large-scale assessment practices and results for students with disabilities.

References

Bielinski, J., Thurlow, M., Minnema, J., & Scott, J. (2000). *How out-of-level testing affects the psychometric quality of test scores*. (Out-of-Level Report 2). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Cleland, W., & Idstein, P. (1980). *In-level versus out-of-level testing of sixth grade special education students*. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Jones, E., Barnette, J., & Callahan, C. (1983, April). *Out-of-level testing for special education students with mild learning handicaps*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Kim, J., & Nicewander, W. (1993). Ability estimation for conventional tests. *Psychometrika, 58*, 587 – 599.

Krueger, R. (1994). *Focus groups: A practical guide for applied research* (2nd Ed.). Thousand Oaks, CA: Sage Publications.

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis* (Out-of-Level Report 1). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Minnema, J., Thurlow, M., & Scott, J. (2001). *Testing students out of level in large-scale assessments: What states perceive and believe* (Out-of-Level Testing Report 5). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

National Center on Educational Outcomes (NCEO). (2001). *Accountability for students with disabilities - NCEO topic area*. Retrieved September 5, 2001, from http://www.coled.umn.edu/nceo/TopicAreas/Accountability/Account_topic.htm

National Center on Educational Outcomes (NCEO). (2001). *FAQ: Universally designed assessments - NCEO topic area*. Retrieved October 16, 2001, from http://www.coled.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesign_FAQ.htm

The Psychological Corporation (1993). *MAT/7 multilevel norms book*: Spring. San Antonio, TX: The Psychological Corporation.

Thompson, S., & Thurlow, M. (2001). *2001 State special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., & Minnema, J. (2001). *States' out-of-level testing policies* (Out-of-Level Testing Report 4). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Yoshida, R. (1976). Out-of-level testing of special education students with a standardized achievement battery. *Journal of Educational Measurement, 13*, 215 – 221.

Appendix A

Focus Group Opening Script

Good afternoon and welcome to our session. Thank you for taking time out of your busy conference schedule to join our discussion on out-of-level testing. My name is Jane Minnema and I am from the National Center on Educational Outcomes. Also with us today from NCEO are Martha Thurlow, John Bielinski, and Dorene Scott.

You were selected because you have knowledge in the area of testing and measurement. We would like to hear your perceptions and opinions about testing students with disabilities out of level in large-scale assessment programs. There are of course no right or wrong answers – but rather different points of view. We welcome both positive and negative opinions. All information will be useful to us.

For the purposes of this discussion, we would like to use the following definition of out-of-level testing proposed by *The Reporting/Accountability Study Group of the Assessing Special Education Students (ASES) State Collaborative on Assessment and Student Standards (SCASS).* You will find this definition on the yellow sheet in your folder.

According to this study group, out-of-level testing is defined as the "administration of a test at a level above or below generally recommended for students based on their age-grade level."

This focus group will last about one and a half hours. Since our time together is limited, we would like to follow up our conversation with an email that will ask one or two more questions. Before we begin, I'd like to share some ground rules. Please speak up – only one person at a time. We are recording our conversation so that we don't miss any of your comments. As you can see, we are on a first name basis this afternoon (evening), but in our reports, no names will be attached to any comments. You may be assured of complete confidentiality. I'd like to begin by asking the first question.

Appendix B

Focus Group Facilitation Protocol

• Background Information – Read opening script.

• Focus Group Questions

Question 1 – Introductory Question

Suppose you have one minute to address this conference about out-of-level testing. What one thing would you say about testing students with disabilities?

Question 2 – Key Question

What are the advantages of using out-of-level testing for students with disabilities in large-scale assessments?

PROBE: When can students with disabilities be tested out of level appropriately?

Question 3 – Key Question

What are the disadvantages of using out-of-level testing for students with disabilities in large-scale assessments?

PROBE: When is out-of-level testing inappropriate for students with disabilities?

Question 4 – Key Question

How can out-of-level testing be used appropriately for system accountability?

PROBE: What is your rationale for that opinion?

Question 5 – Key Question

How can out-of-level testing be used appropriately for student accountability?

PROBE: Again, what is your rationale for that opinion?

• Closing – Present oral summary of responses to the questions.

Question 6 – Ending Question

Is there anything that you would like to add to the summary?

The College of Education
& Human Development

UNIVERSITY OF MINNESOTA

NCEO is an affiliated center of the Institute on Community Integration